Functional preservation of duplicated pair for RSVS III gene in the REST locus of rat 3q42[☆]

Han-Jia Lin[a,*], Chin-Mei Lee[b], Ching-Wei Luo[a], Yee-Hsiung Chen[a,b,*]

[a] *Institute of Biological Chemistry, Academia Sinica, Taipei, Taiwan*
[b] *Institute of Biochemical Sciences, College of Life Science, National Taiwan University, Taipei, Taiwan*

Abstract

Among the major proteins in the rat seminal vesicle secretion, transglutaminase catalyzed the cross-links among RSVS I–III. Six peptide sequences determined from the trypsin digests of RSVS III were confirmed in the protein sequence derived from two paralogs, RSVS III$_\alpha$ and RSVS III$_\beta$ gene, in rat 3q42. Their transcription units are organized with the first exon encoding a signal peptide, and the second a secreted protein, whereas the third encompasses a 3′-non-translated nucleotide that shares common features of rapidly evolving substrates of transglutaminase (REST) gene family. These two genes have 99% identity in their coding region and both express in adult rats with the same transglutaminase cross-linking site, manifesting functional preservation. All of REST genes reported thus far for human and Muridae were mapped in a chromosomal locus between *KCNS1* and *SLPI* suggesting the locus as an active evolving region. The molecular evolution of this gene family is discussed.
© 2004 Elsevier Inc. All rights reserved.

Keywords: Copulatory plug; Gene duplication; Mass spectrum; Rapidly evolving substrates of transglutaminase; Rat; Seminal vesicle

Seminal vesicle is a male mammalian accessory apparatus observed in many but not all mammalian species. Upon ejaculation, seminal vesicle secretions (SVS) stored in the central lumen squirt into urethra to make up the major part of the liquid portion of semen. Since the fertility of mice as well as rat is greatly reduced if their seminal vesicles are extirpated or removed [1,2], this accessory sexual gland has great importance in reproduction. This has prompted studies of the structure and function of protein components of SVS.

The formation of the semi-solid gelatinous mass of the human semen or the seminal plasma clotting in the rodent semen arises subsequent to the discharge of luminal fluid from the male sexual glands and is generally thought to involve transglutaminase-catalyzed cross-linking of proteins in the SVS [3]. In human, the protein substrates have been identified as human semenogelins I and II (HSg I and HSg II), two of the predominant protein components of human SVS [4]. The HSg I/II-like proteins have also been identified in the seminal vesicles of other primates such as the rhesus monkey [5], the cotton-top tamarin [6], and the marmoset [7]. In addition, guinea pig SVP-1 [8], mouse SVS II [9], and rat SVS II [10] in the rodent SVS have all been shown to be good protein substrates of transglutaminase. The genes encoding these protein substrates have similar organization and comprise a characteristic three-exon transcription unit in which the first exon encodes a short signal peptide, the second exon encodes the protein in its entirety, including a termination codon, and the third exon is not translated. Moreover, there is a clear strong similarity between their 5′-flanking regions, first exons, and

* Corresponding authors. Fax: +886 2 23635038.
E-mail addresses: hanjia@gen-info.osaka-u.ac.jp (H.-J. Lin), bc304@gate.sinica.edu.tw (Y.-H. Chen).

3′UTRs, while the protein coding regions in their second exons have experienced rapidly evolutionary change. As a result, the protein substrates in the SVS of different animals show great differences in their primary structure. Accordingly, they are referred to as the rapidly evolving substrates for transglutaminase (REST) gene family [11].

The evolutionary steps undergone by the gene family as well as the physiological significance of their encoded proteins are still obscure. Our previous works have established the genomic structure of mouse SVS III and demonstrated it as a transglutaminase protein substrate involved in copulatory plug formation [12]. Here, we report the duplication of the Muridae SVS III gene in both the rat 3q42 and the mouse 2H3 regions, and the preservation of these paralogous pair with regard to their transglutaminase substrate activity.

Materials and methods

Animals, proteins, and chemicals. Guinea pig liver transglutaminase (EC 2.3.2.13), trypsin, monodansyl-cadaverine, and 3,3′-diaminobenzidine were obtained from Sigma, St. Louis, MO, and α-cyano-4-hydroxycinnamic acid was from Aldrich, Gillingham, England.

Outbred Wistar rats were from the Charles River Laboratories (Wilmington, MA.) and were maintained and bred in the animal center at the College of Medicine, National Taiwan University. The seminal vesicles and coagulating glands of adult rats were carefully dissected free from each other. SVS was collected by squeezing and was diluted with 50 volumes of 50 mM Tris/HCl containing 150 mM NaCl at pH 7.5.

Analytical methods. Protein concentration was determined using the BCA protein assay [13]. According to the method of Laemmli [14], SDS/PAGE was conducted. Assay of the transglutaminase-catalyzed protein cross-links [15] and trypsin digests of protein in the rat SVS followed the previous method [12]. The molecular mass and peptide sequence were determined by a Micromass Q-Tof Ultima MALDI (Waters, Milford, MA). To carry out mass spectrometric analysis, the peptide solution was mixed with an equal volume of the matrix solution (1% α-cyano-4-hydroxycinnamic acid, 50% (v/v) acetonitrile, and 0.1% trifluoroacetic acid) and allowed to air-dry on a sample target. The peptide sequences were identified from the spectral peaks by the Mascot search program (http://www.matrixscience.com/).

Rapid amplification of cDNA ends and genomic cloning. According to the partial cDNA sequence of *pSV-1*, reported by Izawa [16], 5′-rapid amplification of cDNA ends (RACE) was applied to map the transcription initiation site [17]. Total cellular RNA was prepared from the seminal vesicles of adult rats using an Ultraspec-II RNA isolation kit (Biotex, Houston, TX). The 5′-RACE cDNA Synthesis Primer and SMART II oligo (Clontech, Palo Alto, CA) were mixed with 1 μg of freshly prepared rat seminal vesicle total RNA and the first-strand cDNA synthesis was achieved by Superscript II reverse transcriptase (Gibco BRL, Gaithersburg, MD). An oligonucleotide, 5′-GGGG CCTCTTCCTGTTCTTCTCTGTGTCC-3′ which is complementary to nucleotides 167–195 in the RSVS III$_\alpha$ cDNA (Fig. 3A), and Universal Primer Mix were added to the 5′-RACE-ready cDNA to amplify the cDNA fragments which were sequenced to establish the transcriptional initiation site using Advantage 2 Polymerase Mix (Clontech, Palo Alto, CA). Two oligonucleotides, 5′-ACACCACTTTCC

TTTGACTG-3′ and 5′-GAGCCTTCCTGACATCAT-3′, were used to amplify the 3′ ends of RSVS III$_\alpha$ and RSVS III$_\beta$, respectively, by using the same SMART RACE system.

According to the cDNA sequence (Fig. 3A), three oligonucleotides (GSV1-3) were synthesized. GSV1 (5′-CCCGGCTGCTTGCTTCT CCAGAAGGA-3′) and GSV3 (5′-AACACTTGTTTTCTATTCTG TGTAAC-3′) are complementary to nucleotides 57–82 and 1135–1160, respectively. GSV2 (5′-GTGGAGGCCCTTCCTGGTAAG-3′) is complementary to nucleotides 2–27 in the RSVS III$_\alpha$ cDNA (Fig. 3A). PCR was used to amplify DF I from an adaptor-*Dra*I library, DF II/DF III from the *Pvu*II libraries, and DF IV from the *Dra*I library of GenomeWalker kit (Clontech, Pola Alto, CA) using the adaptor primer (AP-1) and each of GSV1-3 as the primer pair (Fig. 3B). The PCR-amplified DNA fragment was then ligated into the pGEM-T-easy vector (Promega, Madison, WI) via TA cloning and the recombinant plasmid was introduced into *Escherichia coli* JM109 strain by transformation.

To identify and quantify the RNA messages of RSVS III$_\alpha$ and RSVS III$_\beta$. Nucleotides 420–435 in the sequence shown in Fig. 3A and another complementary to nucleotides 947–964 were prepared and used as a primer pair to perform RT-PCR with the total RNA (0.2 μg) of rat seminal vesicle. The PCR amplification program was: 94 °C for 30 s, 55 °C for 30 s, 72 °C for 30 s, and cycling 15, 20 or 25 times. The PCR fragment was digested with 0.1 U/μl of *Sph*I at 37 °C for 2 h, and the digested sample was resolved by agarose gel electrophoresis. Each DNA fragment was cloned and sequenced to confirm the expression of both. For quantitative analysis, the gel bands were scanned and analyzed by the computer program, NIH image 1.62 (http://rsb.info.nih.gov/nih-image/Default.html). The intensity of 545 bp gel band was counted for the expression of *RSVS III$_\alpha$*, the 372 bp plus 173 bp bands were counted for *RSVS III$_\beta$*, and reaction products from different amplification cycles were normalized to compare the relative amount of each mRNA.

Northern blotting. A RSVS III$_\alpha$ cDNA fragment (nucleotides 55–435) inserted into the pGEM-T-easy vector or a cDNA fragment of the rat glyceraldehyde-3-phosphated dehydrogenase (GAPDH) gene (nucleotides 1209–1966) inserted into the pGEM3 vector was used as a template to prepare ^{32}P-labeled cDNA probes using a Promega random-priming kit. The general procedures of Northern analysis were followed [18]. The total RNA samples prepared from tissue homogenates were separated on a denaturing 1.5% agarose/formaldehyde gel and then blotted onto a nylon membrane filter by capillary transfer. The membrane was hybridized with one of the two ^{32}P-labeled probes and the RNA messages on the filter were visualized by autoradiography. The probe was then stripped from the membrane and the same membrane was then hybridized with the other ^{32}P-labeled probe. Thus, hybridization with the two probes was performed on the same filter membrane.

The comparison of genomic sequences. The nucleotide sequence of RSVS III$_\alpha$ gene established from this work was mapped on the entire rat genome using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). The cognate REST genes and the two marker genes, *KCNS1* and *SLPI*, were identified by Unigene (http://www.ncbi.nih.gov/entrez/query.fcgi?db=unigene). The chromosome structures of human, mouse, and rat were obtained from Mapviewer (http://www.ncbi.nlm.nih.gov/mapview/). The nucleotide sequences of the two genes were matched by dot plot, using the program, Compare, which was provided by SeqWeb (http://gcg.nhri.org.tw/). The nucleotide sequences of the various genes concerned in this work were collected from EMBL or GenBank and the accession numbers are listed in parentheses for each gene: *HSg I* and *HSg II* (Z47556); rhesus *Sg II* (X92589); tamarin *Sg I* (AJ002153); guinea pig *SVP* (U59711); and mouse *semenoclotin/MSVS II* (XX91270); *MSVS III$_\alpha$* (AF323459); *MSVS III$_\beta$* (NT039210); *MSVS IV* (NM009300); *MSVS V* (NM009301); *MSVS VI* (NM013679); *RSVS II* (J05443); *RSVS III$_\alpha$* (this work); *RSVS III$_\beta$* (XM215936); *RSVS IV* (NM012662); *RSVS V* (NM133516); and *RSVS VI* (XM342575).

Results

Protein characterization of RSVS III

Rat SVS proteins were resolved by SDS/PAGE into five major bands designated RSVS I–V in the decreasing order of $M_r = 94{,}000$–14,000 according to the report of Ostrowski et al. [19] (Fig. 1A, lane 1). RSVS I–III could be cross-linked in reaction buffer by transglutaminase to form a high molecular weight complex, which incorporated monodansyl cadaverine and was retained at the application site of the polyacrylaminde gel (cf. lane 2 of Figs. 1A and B). RSVS I–III were not cross-linked when the enzyme activity was inhibited by the addition of EDTA to the reaction buffer (lanes 1 of Figs. 1A and B).

We digested the 34-kDa RSVS III band from the polyacrylamide with trypsin. The molecular masses of the products were determined by MALDI-TOF from several trypsin digest peaks (a–f on Fig. 2A). The peak e sample was further analyzed to reveal the peptide sequence of GYDLHQDLPQVR (Fig. 2B), which was confirmed as present in the protein derived from the *RpSv-1*, an incomplete gene sequence reported by Izawa [16]. Based on the theoretical molecular mass estimated for each peptide fragment that was predicted from trypsin digestion of the RpSv-1 cDNA-deduced protein sequence,
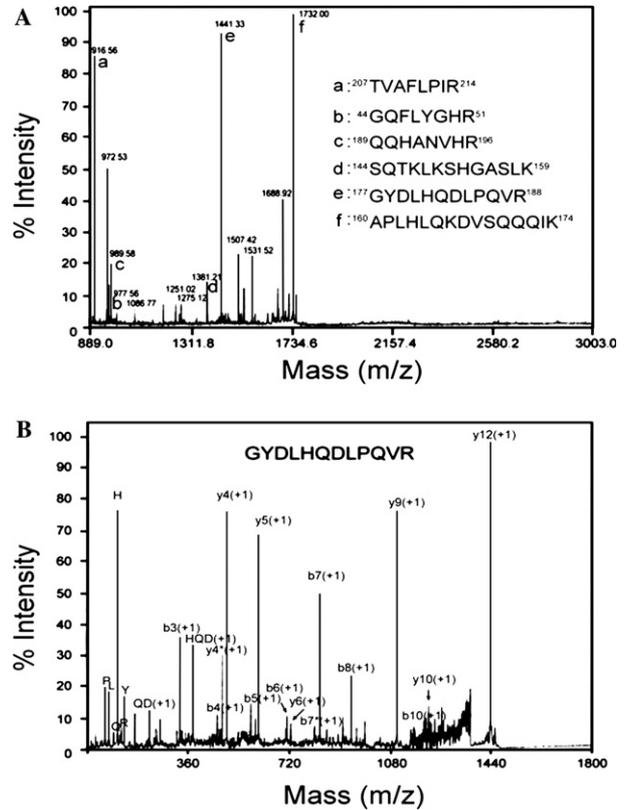
Fig. 2. Determination of the partial sequence in RSVS III. The trypsin-digested RSVS III was subjected to mass spectral analysis. (A) Determination of the molecular mass of each trypsin digest. (B) The peptide sequence of the peak e sample on (A). The peptide sequence listed in (A) was predicted from the Mascot search.

the MS-fit search results matched the peptide sequences of TVAFLPIR for peak a, GQFLYGHR for peak b, QQHANVHR for peak c, SQTKLKSHGASLK for peak d, GYDLHQDLPQVR for peak e, and APLHLQKDVSQQQIK for peak f.

The structural organization of RSVS III gene

According to the RpSv-1 cDNA, we used 5′ and 3′-RACE to establish a complete cDNA sequence of 1184 bp that included a 5′-untranslated region of 22 bp, an open reading frame of 795 bp, which encodes the full length of a polypeptides chain consisting of 264 amino acid residues, and a 3′-untranslated region of 367 bp, which ends with a polyadenylated region (Fig. 3A (GenBank Accession No: AF_525459)). Hereafter, this gene will be referred to as the RSVS III$_\alpha$ gene. Further, we established a genomic sequence consisting of 3986 bp from the four DNA fragments (DF I–IV shown in Fig. 3B). The genomic sequence was completely confirmed in the 3q42 region of rat genome. Alignment of the genomic sequence with the cDNA identified a 5′-flanking region up to −1767 bp from the transcription initiation site, three exons of 95, 1482, and 327 bp, in
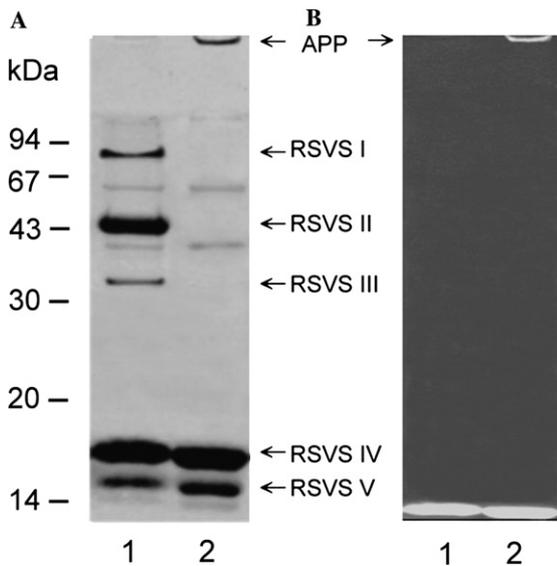
Fig. 1. The transglutaminase-catalyzed protein cross-link among rat SVS protein. To 20 μl of rat SVS solution, 3 μg guinea pig liver transglutaminase in a reaction buffer of 50 mM Tris at pH 7.5, 75 mM Ca^{2+}, and monodansyl cadaverine was added in the absence of (lane 1) or in the presence of 50 mM EDTA (lane 2). After incubation at 37 °C for 60 min, the reaction mixture was resolved by SDS/PAGE using a 14% slab gel ($10 \times 8 \times 0.075$ cm). The gel was stained with Coomassie brilliant blue (A) or observed under ultraviolet light to show the fluorescence arising from monodansyl cadaverine (B). APP is the site of sample application.

A

```
RSVS IIIα  AGTGGAGGCC CTTCCTGGTA AGATGAAGTC CATCTTCTTC AGCCTTTCTC TGCTCCTCCT TCTGGAGAAG CAAGCAGCCG GGATAGGAAT CTATGGTGGG   100
           |||||||||| |||||||||| |||||||||| |||||||||| |||| |||||| |||||||||| |||| ||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ  AGTGGAGGCC CTTCCTGGTA AGATGAAGTC CATCTTCTTC AGCCTTTCTC TGCTCCTCCT TCTGGAGAAG CAAGCAGCCG GGATAGGAAT CTATGGTGGG   100

RSVS IIIα  ACAAAAGGAC ACTTTCTAGT AAAAACATCC CCACTCGTAT TTATCGATAA AGGCCAGTTC CTCTATGGAC ACAGAGAAGA ACAGGAAGAG GCCCCTGAAG   200
           |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ  ACAAAAGGAC ACTTTCTAGT AAAAACATCC CCACTCGTAT TTATCGATAA AGGCCAGTTC CTCTATGGAC ACAGAGAAGA ACAGGAAGAG GCCCCTGAAG   200

RSVS IIIα  AAAGCATCCT TGTTGCAAACT AAGCACCATG TATATAGCCA GGATGCTGAC GCTGACACGG CAGAGGCTCA CGGATCACAA AAGCAGACAG GTTTGAAGGA   300
           |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ  AAAGCATCCT TGTTGCAAACT AAGCACCATG TATATAGCCA GGATGCTGAC GCTGACACGG CAGAGGCTCA CGGATCACAA AAGCAGACAG GTTTGAAGGA   300

RSVS IIIα  AGACATAGTG TGTGATGAAG AAGATGAGCT TGCCCAACAA AAGTCCCAGC TGAAATCCCA GTCACAAATA AAATCCCAAA CCCAAGTAAA ATCCCATGAA   400
           |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ  AGACATAGTG TGTGATGAAG AAGATGAGCT TGCCCAACAA AAGTCCCAGC TGAAATCCCA GTCACAAATA AAATCCCAAA CCCAAGTAAA ATCCCATGAA   400

RSVS IIIα  GCCCAAGTGA AGTCCCAAAC AGGGCAGCTA AAGACCGCGG GGCAGGTGAA ATCACAAACC AAGCTGAAAT CCCACGGAGC TTCTCTGAAG TTCTACAAAG   500
           |||||||||| |||||||||| |||||||||| ||||| || | |||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ  GCCCAAGTGA AGTCCCAAAC AGGGCAGCTA AAGACTGCAG GGCAGGTGAA ATCACAAACC AAGCTGAAAT CCCACGGAGC TTCTCTGAAG TTCTACAAAG   500

RSVS IIIα  CACCACTACA TCTCCAGAAA GACGTTTCTC AGCAGCAAAT CAAGGGCAAA GGATATGACC TGCATCAAGA CCTGCCCCAA GTGCGCCAAC AACATGCAAA   600
           |||||||||| |||||||||| |||||||||| |||||||||| |||||||| | |||||||||| |||||||||| ||| |||||| |||||||||| | ||||||||
RSVS IIIβ  CACCACTACA TCTCCAGAAA GACGTTTCTC AGCAGCAAAT CAAGGGCAGA GGATATGACC TGCATCAAGA CCTTCCCCAA GTGCGCCAAC AGCATGCAAA   600

RSVS IIIα  CGTTCACAGA CTCAAAAGAA AGCTTGGCCA GTCCAGTAAA ACAGTAGCAT TCTTACCCAT TAGACACCGC TTCCAACCCT ATCATGGGTA CTTTACGCAA   700
            ||||||||| |||||||||| |||||||||| |||||||||| ||| |||||| |||||||||| |||||||||| |||||||||| |||||||||| |||| |||||
RSVS IIIβ  TGTTCACAGA CTCAAAAGAA AGCTTGGCCA GTCCAGTAAA ACGGTAGCAT TCTTACCCAT TAGACACCGC TTCCAACCCT ATCATGGGTA CTTTATGCAA   700

RSVS IIIα  TTCCAAGAGC ATCTACATGG CAGTGTTCAT CACACAAAAT CTTTCCCACCA TGGTCCGGGA ATGTGCTACT GTCCAAATGG AGGCTTAATG CTATACCAAG   800
           |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ  TTCCAAGAGC ATCTACATGG CAGTGTTCAT CACACAAAAT CTTTCCCACCA TGGTCCGGGA ATGTGCTACT GTCCAAATGG AGGCTTAATG CTATACCAAG   800

RSVS IIIα  GCATCTTCAC TGAATAACAC ATTCGTTAAT CTCTCAAGAA TTAAAACACG TGTAGAGgag tcacctgacc tgtgtggact gtgcggttct taagatgctg   900
           |||||||||| ||| ||| || |||| |||| | |||||||||| |||||||||| |||||||||||||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ  GCACCTTCAC TGAGTAACAC ATTCATTAAT CTCTCAAGAA TTAAAACACG TGTAGAGgag tcacctgacc tgtgtggact gtgcggttct taagatgctg   900

RSVS IIIα  actgccgtgt ggtcctcaac accttgcacg gatgccatga cccactcaca cattccttcc ..agagactt tctaaagttg gagcaacctc acgcaggtgc   998
           ||||| ||| |||||||||| |||||||||| |||||||||| |||||||||| |||||||||   |||||||| |||||||||| ||||||| || ||||||||||
RSVS IIIβ  actgccctgt ggtcctcaac accttgcacg gatgccatga cccactcaca cattccttcc agagagactt tctaaagttg gagcaacttc acgcaggtgc   1000

RSVS IIIα  cctccaataa acagagcctt cccgacacca ctttcctttg actgtcaaaa tctccaagtt cctattggct ttggtgcttg gatggcatct tttcaagtaa   1098
           |||||||||| |||||||||| || ||||| || || | ||| |         | | ||| |  |||       ||| | ||| ||| | ||| ||| ||
RSVS IIIβ  cctccaataa acagagcctt cctgacatca tttgcttttt taaaaaatgt ttttcatttg aaatagaatc atatcccttt ccctgttttt cttctcttca   1100

RSVS IIIα  taaactcaaa ttgaaagaac agaacactct tgtagagtta cacagaaatag aaaacaagtg tttcccaata aagacattgt tttggc              1184
              |||                |||       | ||      |        |  | ||    |        | || |         | || |||
RSVS IIIβ  acccctcccc agcatccttt tctgtcttct ccacagttcc cagctctgaa gttcgtaaca tctctgtttt tgagtgtcgt ttttaaatac ctatatatgt   1200

RSVS IIIβ  atgtatattc ataaatatat aagtacaact tgctg                                                                        1234
```

B
```
        -1767                        +1  95   324                         1805 1893  2219
         ┌─────────────────────────┬─■─┬───■────────────────────────────■──┬──■───┬────┐
         └─────────────────────────┴───┴────────────────────────────────┴──────┴────┘
   AP1 →
         ─────────────────────────────────────  ← GSV1
                    (DF I)           GSV2 →
                                        ──────  ← AP1
                                       (DF II)
                                    AP1 →
                                        ────────────────────── ← GSV3
                                              (DF III)
                              GSV2 →
                                    ────────────────────────── ← AP1
                                              (DF IV)
```

Fig. 3. The genomic structures responsible for RSVS III. The RSVS IIIα cDNA sequence consisting of 1184 bp was established by cDNA cloning. The nucleotides of exon 1 and 2 are in capital letters and those downstream exon 2, indicating exon 3 and partial 3′-UTR, are in lowercase letters. RSVS IIIβ cDNA sequence was mapped from the rat genomic sequence and conformed with RACE analysis. Nucleotide numbers are relative to the transcription start marked by ∗. The translation start and stop codons and the polyadenylated signal are marked by solid lines. The SphI restriction site in RSVS IIIβ cDNA is denoted by an open box. (B) The nucleotide sequences of DF I–IV (see Materials and methods) in a schematic diagram were aligned to construct the sequence consisting of 3986 nucleotides, which were confirmed in the RSVS IIIα genomic sequence from the entire genomic sequence of rat. The genomic structure includes three exons denoted by E1–3 (black bars), which were established from the cDNA sequence shown in (A), 5′-flanking region, and 2 introns (open bars).

which the stop codon appears in the second exon, and two introns of 230 and 89 bp (Fig. 3B). Thus, the transcription unit of RSVS IIIα gene shows all the characteristics of a REST gene.

In rat chromosome 3q42 region, we also mapped another gene designated *RSVS IIIβ* which shares a very high similarity with *RSVS IIIα* through the entire geno-mic sequence from 1.5 kb upstream of exon 1 to part of exon 3. According to the dot plot match for the two nucleotide sequences, 100%, 99.6%, 98.6%, and 94.8% of nucleotides in the exon 1, intron 1, exon 2, and intron 2 regions are identical, respectively. By using 5′ and 3′-RACE, we established the full length cDNA of RSVS IIIβ (Fig. 3A). There is a certain degree of variation in

the nucleotide sequence downstream from the second intron such that the nucleotide identity declines to 60% in exon 3. The six peptide sequences established from the mass spectral analysis were confirmed in both proteins derived from these two sibling genes. If both are translated, the two proteins contain an equal number of amino acid residues with almost identical amino acid sequences at every position, except that K^{177}, T^{227}, and I^{263} of RSVS III$_\alpha$ are replaced by R^{177}, M^{227}, and T^{263} in RSVS III$_\beta$ (Fig. 4). Furthermore, we searched the mouse genome for the cognate genes. Surprisingly, a similar situation occurs to the genes encoding MSVS III$_\alpha$ and MSVS III$_\beta$, only N^{45}, G^{82}, V^{143}, and F^{159} of

MSVS III$_\alpha$ are replaced by S^{45}, E^{82}, M^{143}, and Y^{159} of MSVS III$_\beta$ (Fig. 4). It is worth noting that identity between two rat proteins is higher than any protein from mouse and vice versa.

The exclusive expression of RSVS III$_\alpha$/RSVS III$_\beta$ in seminal vesicles

The RSVS III$_\beta$ cDNA contains ^{592}GCATGC^{596} that is an *Sph*I restriction site and this sequence is replaced by ^{592}ACATGC^{596} in RSVS III$_\alpha$ cDNA (Fig. 3A). This provides a method by which RSVS III$_\beta$ and RSVS III$_\alpha$ mRNA can be distinguished. As shown in Fig. 5A (in-


```
RSVS IIIα   MKSIFFSLSL LLLLEKQAAG IGIY.GGTKG HFLVKTSPLV FIDKGQFLYG   49
            |||||||||| |||||||||| |||| ||||| |||||||||| ||||||||||
RSVS IIIβ   MKSIFFSLSL LLLLEKQAAG IGIY.GGTKG HFLVKTSPLV FIDKGQFLYG   49
            |||||||||| |||||| ||| |  | ||||| |||||||||| ||| |||||
MSVS IIIα   MKSIFFSLSL LLLLEKKAAG IELYAGGTKG HFLVKTSPLM FIGKNQFLYG   50
            |||||||||| |||||||||| |||||||||| |||||||||| |||||||||
MSVS IIIβ   MKSIFFSLSL LLLLEKKAAG IELYAGGTKG HFLVKTSPLM FIGKSQFLYG   50

RSVS IIIα   HREEQEEAPE ESILVQTKHH VYSQDADADT AEAHGSQKQT GLKEDIVCDE   99
            |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ   HREEQEEAPE ESILVQTKHH VYSQDADADT AEAHGSQKQT GLKEDIVCDE   99
            | |||||||| ||| ||||||| |  ||||||  || ||  | |||||| |
MSVS IIIα   HKEEQEEAPE ESIFVQTKHH AYGQDADADM GGALSSQELT SLKEDIVCEE  100
            |||||||||| |||||||||| |||||||||| | |||||||| ||||||||||
MSVS IIIβ   HKEEQEEAPE ESIFVQTKHH AYGQDADADM GEALSSQELT SLKEDIVCEE  100

RSVS IIIα   EDELAQQKSQ LKSQSQIKSQ TQVKSHEAQV KSQTGQLKTA GQVKSQTKLK  149
            |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ   EDELAQQKSQ LKSQSQIKSQ TQVKSHEAQV KSQTGQLKTA GQVKSQTKLK  149
            |||||||||| | || ||||| |||| || |  ||| |||| | || ||| ||
MSVS IIIα   EDELAQQKSQ LPSQSQIKSQ TQVKSYAAQL KSQPGQLKTI GQVKSQTMLK  150
            |||||||||| |||||||||| |||| |||| | ||| ||||| || ||||||
MSVS IIIβ   EDELAQQKSQ LPSQSQIKSQ TQVKSYAAQL KSQPGQLKTI GQMKSQTMLK  150

RSVS IIIα   SHGASLKFYK APLHLQKDVS QQQIKGKGYD LHQDLPQVRQ QHANVHRLKR  199
            |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ   SHGASLKFYK APLHLQKDVS QQQIKGRGYD LHQDLPQVRQ QHANVHRLKR  199
            |||| || || | | | || | || ||||| | || |||| | | |||||
MSVS IIIα   SHGAPLKSFK ARLNLREDI. PQQVKGRGYG LAEDLAQVRQ QPAKVHRLKG  199
            |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
MSVS IIIβ   SHGAPLKSYK ARLNLREDI. PQQVKGRGYG LAEDLAQVRQ QPAKVHRLKG  199

RSVS IIIα   KLGQSSKTVA FLP.IRHRFQ PYHGYFTQFQ EHLHGSVHHT KSFHHGPGMC  248
            |||||||||| ||| ||||||| |||||||||| |||||||||| ||||||||||
RSVS IIIβ   KLGQSSKTVA FLP.IRHRFQ PYHGYFMQFQ EHLHGSVHHT KSFHHGPGMC  248
            | || || | | | || || || | ||| | | ||||| ||| ||||
MSVS IIIα   KHRQSRKTAA FYPQFRRRSR PYPRYFVQFQ EQLQGSVHHT KSFYPGPGMC  249
            |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
MSVS IIIβ   KHRQSRKTAA FYPQFRRRSR PYPRYFVQFQ EQLQGSVHHT KSFYPGPGMC  249

RSVS IIIα   YCPNGGLMLY QGTFTE                                       264
            |||||||||| || |||
RSVS IIIβ   YCPNGGLMLY QGTFTE                                       264
            ||| || || | ||
MSVS IIIα   YCPRGGVILY QDAFTD                                       265
            |||||||||| ||||||
MSVS IIIβ   YCPRGGVILY QDAFTD                                       265
```

Fig. 4. The homologous alignment of the Muridae SVS III proteins. The protein sequences deduced from the four genes responsible for RSVS III$_\alpha$, RSVS III$_\beta$, MSVS III$_\alpha$, and MSVS III$_\beta$ were aligned from the translational start. Identical sequence is denoted by a vertical bar. The variant amino acids in the pair of RSVS III$_\alpha$/RSVS III$_\beta$ or MSVS III$_\alpha$/MSVS III$_\beta$ are shown by gray or black boxes. The peptide sequences determined from mass spectral analysis for the trypsin digests of RSVS III are indicated by solid lines. The tandem repeats of the short peptide sequence QXK(S/T) are denoted in hatched boxes.

set), the PCR product for nucleotides 420–964 in Fig. 3A was a single 545 bp band and, on *Sph*I digestion, it was resolved to three bands of 545, 372, and 173 bp by agarose gel electrophoresis. The 545-bp fragment was confirmed to be nucleotides 420–964 of RSVS III$_\alpha$ cDNA, and the 173-bp fragment and 372-bp fragment were, respectively, confirmed to be the nucleotides 420–592 and the nucleotides 593–964 of RSVS III$_\beta$ cDNA. This supports the expression of both genes in the seminal vesicle. Because the PCR primer pairs we used have the same affinity to both genes, the PCR product of each gene from the same application cycle should be proportional to the amount of the template before the reaction reached the plateau. Based on densitometric scanning of the results from different amplification cycles, RSVS III$_\alpha$ mRNA was about twofold higher than that of the RSVS III$_\beta$ mRNA (Fig. 5A).

We used the cDNA fragment, nucleotides 55–435, which appear in both RSVS III$_\alpha$ and RSVS III$_\beta$ cDNA, to probe the total RNAs prepared from the reproductive tracts, including seminal vesicle, epididymis, testis, coagulating gland, vas deferens, prostate, uterus, vagina, ovary, as well as non-reproductive organs such as lung,

kidney, brain, spleen, liver, pancreas, and heart. The RNA message was abundant only in seminal vesicle and was not detectable in the other tissues, indicating an exclusive expression of both of these genes in seminal vesicle (Fig. 5B).

Homology search of REST genes

Except *RSVS I*, we found that the genes encoding the major rat SVS proteins are clustered at a 270 kb region between a potassium voltage-gated channel, *KCNS1* [20,21], and a secretory leukocyte protease inhibitor, *SLPI* [22], on the rat 3q42 by the downstream order of *RSVS II*, *RSVS III$_\beta$*, *RSVS IV*, *RSVS III$_\alpha$*, *RSVS VI*, and *RSVS V* with the first two having the same gene orientation and the final four having an opposed orientation (Fig. 6A). Despite the fact that RSVS IV and RSVS V are not involved in the transglutaminase-catalyzed protein cross-links in the rat SVS (Fig. 1A), their genomic sequences still show the characteristic three exon–two intron structure of a REST gene [11]. RSVS VI is not involved in the plug formation either. Although it is encoded by a four-exon genomic struc-
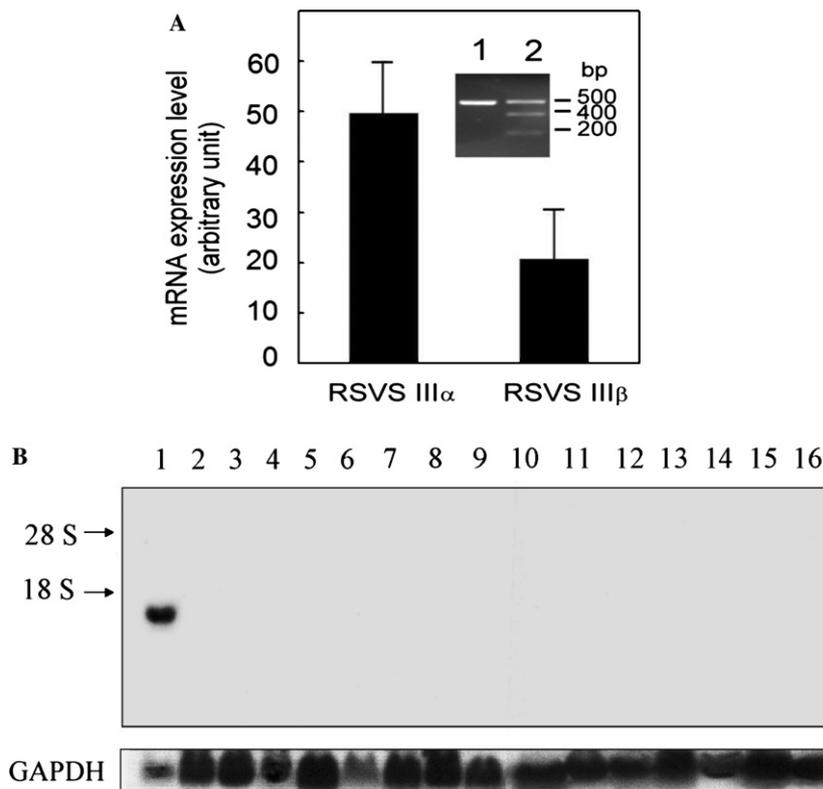
Fig. 5. Expression of the mRNAs for RSVS III$_\alpha$ and RSVS III$_\beta$. (A) A DNA fragment of nucleotides 420–964 in both RSVS III$_\alpha$ cDNA and RSVS III$_\beta$ cDNA was amplified by PCR. The amplified DNA alone (lane 1, inset) or its *Sph*I-digested sample (lane 2, inset) was resolved by electrophoresis on a 1% agarose gel, which was stained with ethidium bromide to show the DNA bands. The intensity corresponding to RSVS III$_\alpha$ and RSVS III$_\beta$ mRNA was measured by three experiments. The bars represent SD after normalization. (B) Northern analysis of the total RNA from seminal vesicle (lane 1), epididymis (lane 2), testis (lane 3), coagulating gland (lane 4), vas deferens (lane 5), prostate (lane 6), uterus (lane 7), vagina (lane 8), ovary (lane 9), lung (lane 10), kidney (lane 11), brain (lane 12), spleen (lane 13), liver (lane 14), pancreas (lane 15), and heart (lane 16), using a RSVS III cDNA probe.
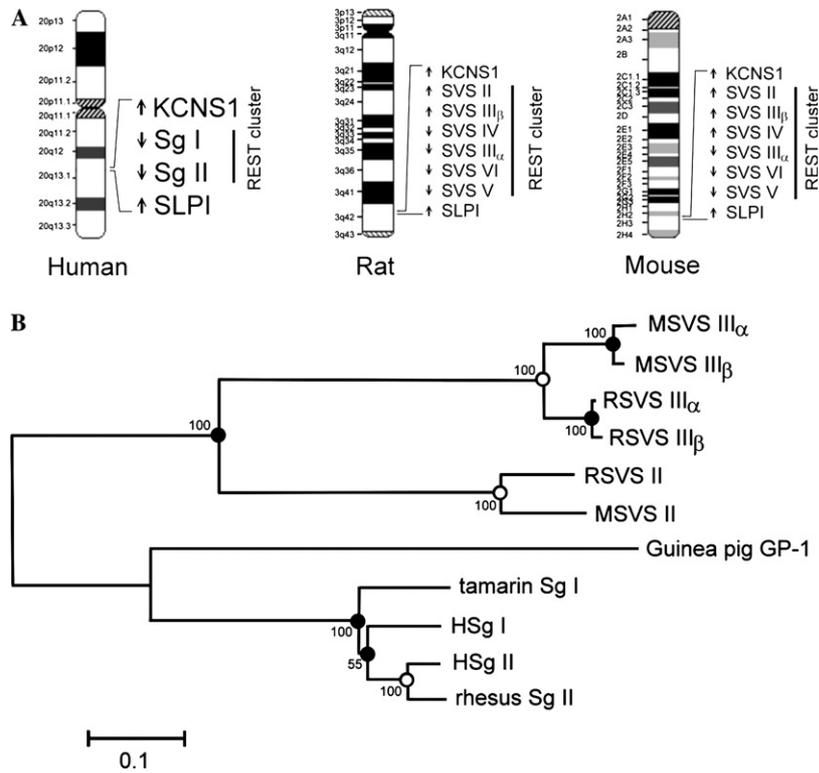
Fig. 6. A phylogenic tree of REST gene family. (A) The location of REST genes in the Muridae and human chromosomes. (B) The construction of a phylogenic tree by the neighbor-joining method. The scale bar corresponds to 0.10 substitution per site and the bootstrap percentage on the basis of 1000 replicates is given at each duplication (●) and speciation (○).

ture, its exons 1 and 4 show similar characteristics to exons 1 and 3 of a REST gene [23]. Based on the aforementioned features, we referred to this chromosomal region as the REST locus. Using *KCNS1* and *SLPI* as landmarks, we were able to define a similar 200 kb mouse REST locus made up of *MSVS II*, *MSVS III_β*, *MSVS IV*, *MSVS III_α*, *MSVS VI*, and *MSVS V* at the mouse 2H3 locus with a similar order of genes as their cognates in the rat 3q42 locus (Fig. 6A). Likewise, the *HSg I* and *HSg II* were found within the 260 kb region between *KCNS1* and *SLPI* at the human 20q13 locus (Fig. 6A). Thus, the REST genes exclusively expressed in seminal vesicles may be clustered between *KCNS1* and *SLPI* on a mammalian chromosome.

To clarify the evolutional relationship of REST genes from different mammals, we applied the neighbor-joining method, the maximum parsimony method, the minimum evolution method, and the unweighted pair-group method in Data Analysis in Molecular Biology and Evolution (DAMBE) [24] to construct the phylogenic tree from the 11 known REST genes whose translated proteins are involved in transglutaminase-catalyzed cross-linking in SVS. The phylogenetic tree from each method shows a similar pattern. Fig. 6B displays one representative pattern established from the neighbor-joining method. These 11 genes are likely to have a common evolutionary origin, and they are divided to two sub-

groups, with the Muridae genes in one subgroup, and *GP1G* and the primate genes in the other. The nucleotide substitution frequency suggested that *GP1G* and the Muridae genes evolved faster than the primate genes in which the human genes slowly evolved relative to the tamarin/rhesus gene. It is of interest to note that *GP1G* underwent rapid evolution without gene duplication after it diverged from the primate semenogelin progenitor. The primate REST ancestor gene duplicated to generate the two human paralogs, one *HSg I* ortholog in tamarin, and one *HSg II* ortholog in rhesus monkey. However, the Muridae REST genes experienced gene duplication twice. Despite the relatively high substitution rate in Muridae REST gene, the highly homologous *RSVS III_α/RSVS III_β* and *MSVS III_α/MSVS III_β* pairs suggesting these genes were duplicated recently. We also noted that the nucleotide substitution rate for the three rat genes seems to be slower than the cognate mouse genes, implying that rat may be closer to the Muridae ancestor and this agrees with the fossil records [25].

Discussion

Except *RSVS I/MSVS I*, the genes encoding the major Muridae SVS proteins are assembled in one chromosomal region designated the REST locus. Their genomic

structures preserve a characteristic three exon–two intron structure of the REST gene. A high level of sequence identity in their 5′-flanking regions and first exons suggests convergent evolution is necessary to effect a high-level expression of these secretory proteins in the seminal vesicle, which has evolved to become the only tissue capable of REST gene expression. The unusual divergence of their second exons seems to be coupled with the development of new gene functions. Apparently, the mammalian REST locus has been a very actively evolving region.

The SVS proteins may be functionally divided into two groups, one with the transglutaminase substrate activity for the semen gelating or clotting and the other group having another function lack of the clotting ability. The transglutaminase protein substrates derived from REST genes have a characteristically high content of glutamine and lysine residues; both residues make up 22%, 27%, 30%, 20%, 21%, and 20% of the total amino acid residues in HSg I, MSVS II, RSVS II, MSVS III$_\alpha$/MSVS III$_\beta$, RSVS III$_\alpha$/RSVS III$_\beta$, and guinea pig SVP-1, respectively. These two kinds of amino acids are frequently localized in a specific peptide sequence that is tandem-repeated in the central part of each protein. For example, there are 20 tandem repeats of the sequence QVKSSGS that extends from residue 82 to 250 in MSVS II that consists of 375 amino acid residues [9]. RSVS II contains 414 amino acid residues in which there are 13 repeats of a consensus sequence QSQLKSFGQVKSS spanning residues 86–298 of the protein molecule [10]. A short peptide QXK(S/T) in which X represents an aliphatic amino acid residue is tandem-repeated 5 times in residues 116–145 of MSVS III$_\alpha$/MSVS III$_\beta$ [12]. Such an oligopeptide sequence is tandem-repeated 6 times in the central part of both RSVS III$_\alpha$ and RSVS III$_\beta$ (Fig. 4, residues 109–144).

It is believed that the repeat segments in these Muridae RESTs result from internal duplication and that enhance its participation in the plug formation [11]. Previously, we demonstrated that one short peptide of QXK(S/T) in MSVS III is sufficient to act as a transglutaminase substrate acyl donor as well as an acyl acceptor for enzyme cross-linking [12]. Since RSVS III$_\alpha$ and RSVS III$_\beta$ contain equal number of such an oligopeptide substrate sequence, they may have the same transglutaminase substrate activity. Together with the finding that the RNA messages of both RSVS III$_\alpha$ and RSVS III$_\beta$ are present in the seminal vesicle, this suggests the functional preservation of these paralogous genes. Apparently, no member of the paralogous pairs has degenerated to become a pseudogene. This also suggests that the transcriptional elements, controlling the specific expression manner of *RSVS III*, might be preserved in the duplication segment.

The peptide repeats in HSg and guinea pig SVP-1 are different from those in the Muridae proteins.

From sequence analysis, the second exons of the genome in a REST locus can be rapidly diversified after the mammalian speciation by means such as internal repetition or frequent mutation This has led to variations in length and the sequence of tandem-repeated segments in each REST gene. Moreover, Hagstrom et al. [27] suggested that GP1G and HSg II gene evolved from a common ancestral gene with a single sequence, which has been selected for function as a coding exon in *HSg II* and as the first intron in *GP1G*. Such finding provides an explanation for the mechanisms of the rapid evolution of REST genes. The rapid evolving of exon 2 might give species-specific transglutaminase-catalyzed sites or an active site for other functions.

On the other hand, no tandem-repeated sequences appear in MSVS IV/RSVS IV, MSVS V/RSVS V, and MSVS VI/RSVS VI. These proteins have no clotting ability and only exist in Muridae REST locus. With more genes evolved (Fig. 6A) and higher substitution rate than the guinea pig and primate genes (Fig. 6B), the REST locus of Muridae is even more active than other mammalians. These specialized REST genes in Muridae may play roles other than plug formation to meet their physiological needs. For example, RSVS IV has been considered to be an immunomodulator [26].

It is of interest to note that the REST genes with clotting abilities were also reported to have other functions. For instance, RSVS II is present in the nuclear matrix of seminal vesicle to bind the promoter of RSVS IV gene [28]. HSg I is able to inhibit sperm motility in addition to its transglutaminase substrate activity [29]. *GP1G* encodes a precursor protein that is cleaved to three proteins with a different function in a post-translational proteolytic process [30]. Taken together, the semen gelating or clotting ability may be not the only selective pressure that has prompted the REST genes to undergo rapid evolution in order to acquire different physiological roles.

This work is the first to demonstrate that the sibling genes *RSVS III$_\alpha$* and *RSVS III$_\beta$* together with *MSVS III$_\alpha$* and *MSVS III$_\beta$* are apparently the last duplication of a paralogous pair in the Muridae REST locus. It is of interest to note that the second exons of the paralogous pair are highly conserved and there is preservation of their function. Thus, duplication of the Muridae paralogs does not fit the classical model for the evolution of duplicated genes, but is consistent with the new conceptual framework of Force et al. [31,32], who have shown how complementary degenerative mutations in different regulatory elements of duplicated genes can facilitate the preservation of both duplicates. The paralogs of Muridae *SVS III* may serve as "living models" in the future to trace the molecular evolution of the REST gene family.

References

[1] S.F. Pang, P.H. Chow, T.M. Wong, The role of the seminal vesicles, coagulating glands and prostate glands on the fertility and fecundity of mice, J. Reprod. Fertil. 56 (1979) 129–132.

[2] B. Peitz, P. Olds-Clarke, Effects of seminal vesicle removal on fertility and uterine sperm motility in the house mouse, Biol. Reprod. 35 (1986) 608–617.

[3] H.G. Williams-Ashman, Transglutaminases and the clotting of mammalian seminal fluids, Mol. Cell. Biochem. 58 (1984) 51–61.

[4] A. Peter, H. Lilja, A. Lundwall, J. Malm, Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase, Eur. J. Biochem. 252 (1998) 216–221.

[5] M. Ulvsback, A. Lundwall, Cloning of the semenogelin II gene of the rhesus monkey. Duplications of 360 bp extend the coding region in man, rhesus monkey and baboon, Eur. J. Biochem. 245 (1997) 25–31.

[6] A. Lundwall, The cotton-top tamarin carries an extended semenogelin I gene but no semenogelin II gene, Eur. J. Biochem. 255 (1998) 45–51.

[7] A.E. Huber, H. Fraser, J. Rio-Garcia, G. Kreil, D.G. Smyth, Molecular cloning in the marmoset shows that semenogelin is not the precursor of the TRH-like peptide pGlu-Glu-Pro amide, Biochim. Biophys. Acta 1387 (1998) 143–152.

[8] J.T. Moore, J. Hagstrom, D.J. McCormick, S. Harvey, B. Madden, E. Holicky, D.R. Stanford, E.D. Wieben, The major clotting protein from guinea pig seminal vesicle contains eight repeats of a 24-amino acid domain, Proc. Natl. Acad. Sci. USA 84 (1987) 6712–6714.

[9] A. Lundwall, The cloning of a rapidly evolving seminal-vesicle-transcribed gene encoding the major clot-forming protein of mouse semen, Eur. J. Biochem. 235 (1996) 424–430.

[10] S.E. Harris, M.A. Harris, C.M. Johnson, M.F. Bean, J.G. Dodd, R.J. Matusik, S.A. Carr, J.W. Crabb, Structural characterization of the rat seminal vesicle secretion II protein and gene, J. Biol. Chem. 265 (1990) 9896–9903.

[11] A. Lundwall, C. Lazure, A novel gene family encoding proteins with highly differing structure because of a rapidly evolving exon, FEBS Lett. 374 (1995) 53–56.

[12] H.J. Lin, C.W. Luo, Y.H. Chen, Localization of the transglutaminase cross-linking site in SVS III, a novel glycoprotein secreted from mouse seminal vesicle, J. Biol. Chem. 277 (2002) 3632–3639.

[13] P.K. Smith, R.I. Krohn, G.T. Hermanson, A.K. Mallia, F.H. Gartner, M.D. Provenzano, E.K. Fujimoto, N.M. Goeke, B.J. Olson, D.C. Klenk, Measurement of protein using bicinchoninic acid, Anal. Biochem. 150 (1985) 76–85.

[14] U.K. Laemmli, Cleavage of structural proteins during the assembly of the head of bacteriophage T4, Nature 227 (1970) 680–685.

[15] A. Lundwall, A. Peter, J. Lovgren, H. Lilja, J. Malm, Chemical characterization of the predominant proteins secreted by mouse seminal vesicles, Eur. J. Biochem. 249 (1997) 39–44.

[16] M. Izawa, cDNA cloning of androgen-stimulated mRNAs in rat seminal vesicles: partial characterization of newly isolated cDNA clones, pSv-1 and pSv-2, Endocrinol. Jpn. 37 (1990) 223–232.

[17] B. Gong, R. Ge, Using the SMART cDNA system to map the transcription initiation site, Biotechniques 28 (2000) 846–848, 850, 852.

[18] J. Sambrook, E.F. Fritsch, T. Maniatis, Molecular Cloning: A Laboratory Manual, second ed., Cold Spring Harbor, 1989, pp. 7.19–7.50 and p. 9.58.

[19] M.C. Ostrowski, M.K. Kistler, W.S. Kistler, Purification and cell-free synthesis of a major protein from rat seminal vesicle secretion. A potential marker for androgen action, J. Biol. Chem. 254 (1979) 383–390.

[20] M. Salinas, F. Duprat, C. Heurteaux, J.P. Hugnot, M. Lazdunski, New modulatory alpha subunits for mammalian Shab K$^+$ channels, J. Biol. Chem. 272 (1997) 24371–24379.

[21] M. Stocker, D. Kerschensteiner, Cloning and tissue distribution of two new potassium channel alpha-subunits from rat brain, Biochem. Biophys. Res. Commun. 248 (1998) 927–934.

[22] R.C. Thompson, K. Ohlsson, Isolation, properties, and complete amino acid sequence of human secretory leukocyte protease inhibitor, a potent inhibitor of leukocyte elastase, Proc. Natl. Acad. Sci. USA 83 (1986) 6692–6696.

[23] A.M. Simon, G. Veyssiere, C. Jean, Structure and sequence of a mouse gene encoding an androgen-regulated protein: a new member of the seminal vesicle secretory protein family, J. Mol. Endocrinol. 15 (1995) 305–316.

[24] X. Xia, Z. Xie, DAMBE: software package for data analysis in molecular biology and evolution, J. Hered. 92 (2001) 371–373.

[25] L. Jacobs, D. Pilbeam, Of mice and men: fossil-based divergence dates and molecular "clock", J. Hum. Evol. 9 (1980) 551–555.

[26] G. Peluso, R. Porta, C. Esposito, M.A. Tufano, R. Toraldo, M.L. Vuotto, G. Ravagnan, S. Metafora, Suppression of rat epididymal sperm immunogenicity by a seminal vesicle secretory protein and transglutaminase both in vivo and in vitro, Biol. Reprod. 50 (1994) 593–602.

[27] J.E. Hagstrom, M.P. Fautsch, M. Perdok, A. Vrabel, E.D. Wieben, Exons lost and found. Unusual evolution of a seminal vesicle transglutaminase substrate, J. Biol. Chem. 271 (1996) 21114–21119.

[28] M.J. Horton, R.H. Getzenberg, Rat seminal-vesicle secretory protein SVS II binds DNA with a preference for the 5′ regulatory region of secretory protein SVS IV gene: co-isolation with components of the nuclear matrix, J. Androl. 20 (1999) 267–279.

[29] M. Robert, C. Gagnon, Purification and characterization of the active precursor of a human sperm motility inhibitor secreted by the seminal vesicles: identity with semenogelin, Biol. Reprod. 55 (1996) 813–821.

[30] J.T. Moore, M.E. Norvitch, C.M. Veneziale, The cDNA cloning of a 5 5-kilodalton protein from guinea pig seminal vesicle. Evidence that the protein is the precursor of a 25-kilodalton basic secretory protein, J. Biol. Chem. 260 (1985) 3826–3832.

[31] M. Lynch, A. Force, The probability of duplicate gene preservation by subfunctionalization, Genetics 154 (2000) 459–473.

[32] A. Force, M. Lynch, F.B. Pickett, A. Amores, Y.L. Yan, J. Postlethwait, Preservation of duplicate genes by complementary, degenerate mutations, Genetics 151 (1999) 1531–1545.